D. Requirements for Additional Computing Resources

We believe that one additional D-machine would improve our computing efficiency. Experience indicates that one machine per person is desirable, allowing each person to preserve his work and begin again without delay.

E. Recommendations for Future Community and Resource Development

All GUIDON and NEOMYCIN work has shifted shifting to Xerox Dolphins and Dandelions (D-machines), with the exception of daily mail and text editing, which is performed on the DEC 2060. Given that D-machines cannot be accessed remotely, work at home is restricted to what can be done on the 2060. We believe that the availability of personal workstations (the D-machines) lessens the need to do programming at home. However, we still find it convenient to be able to write at home. Integration of the workstation graphics with our wordprocessor, made available at home, would greatly increase our ability to write and prepare presentations.

Difficulties with the current resource include: unreliable file server for D-machines and inadequate utilities for examining a directory and deleting files. While the situation has greatly improved over the past few years, basic development is still required. Given the amount of software delivered in the recent past, the delay in producing an integrated environment is understandable. It is clear that a central resource group will be necessary for some time to come, charged with integrating networks, file servers, and attached devices to respond to individual requests.

IV.A.2. MOLGEN Project

MOLGEN - Applications of Artificial Intelligence to Molecular Biology: Research in Theory Formation, Testing, and Modification

Prof. E. Feigenbaum and Dr. P. Friedland
Department of Computer Science
Stanford University

Prof. Charles Yanofsky
Department of Biology
Stanford University

I. SUMMARY OF RESEARCH PROGRAM

A. Project Rationale

The MOLGEN project has focused on research into the applications of symbolic computation and inference to the field of molecular biology. This has taken the specific form of systems which provide assistance to the experimental scientist in various tasks, the most important of which have been the design of complex experiment plans and the analysis of nucleic acid sequences. Our current research concentrates on scientific discovery within the subdomain of regulatory genetics. We desire to explore the methodologies scientists use to modify, extend, and test theories of genetic regulation, and then emulate that process within a computational system.

Theory or model formation is a fundamental part of scientific research. Scientists both use and form such models dynamically. They are used to predict results (and therefore to suggest experiments to test the model) and also to explain experimental results. Models are extended and revised both as a result of logical conclusions from existing premises and as a result of new experimental evidence.

Theory formation is a difficult cognitive task, and one in which there is substantial scope for intelligent computational assistance. Our research is toward building a system which can form theories to explain experimental evidence, can interact with a scientist to help to suggest experiments to discriminate among competing hypotheses, and can then revise and extend the growing model based upon the results of the experiments.

The MOLGEN project has continuing computer science goals of exploring issues of knowledge representation, problem-solving, discovery, and planning within a real and complex domain. The project operates in a framework of collaboration between the Heuristic Programming Project (HPP) in the Computer Science Department and various domain experts in the departments of Biochemistry, Medicine, and Biology. It draws from the experience of several other projects in the HPP which deal with applications of artificial intelligence to medicine, organic chemistry, and engineering.

B. Medical Relevance and Collaboration

The field of molecular biology is nearing the point where the results of current research will have immediate and important application to the pharmaceutical and chemical industries. Already, clinical testing has begun with synthetic interferon and human growth hormone produced by recombinant DNA technology. Governmental reports estimate that there are more than two hundred new and established industrial firms already undertaking product development using these new genetic tools.

The programs being developed in the MOLGEN project have already proven useful and important to a considerable number of molecular biologists. Currently several dozen researchers in various laboratories at Stanford (Prof. Paul Berg's, Prof. Stanley Cohen's, Prof. Laurence Kedes', Prof. Douglas Brutlag's, Prof. Henry Kaplan's, and Prof. Douglas Wallace's) and over four hundred others throughout the country have used MOLGEN programs over the SUMEX-AIM facility. We have exported some of our programs to users outside the range of our computer network (University of Geneva [Switzerland], Imperial Cancer Research Fund [England], and European Molecular Biology Institute [Heidelberg] are examples). The pioneering work on SUMEX has led to the establishment of a separate NIH-supported facility, BIONET, to serve the academic molecular biology research community with MOLGEN-like software. BIONET is now serving many of the computational needs of over two thousand academic molecular biologists in the United States.

C. Highlights of Research Progress

C.1 Accomplishments

During the past year we have concentrated on the qualitative modeling and simulation aspects of the research. Our view is that a well-formulated, multi-level model of a scientific theory is a necessary first step to automated discovery. In addition, we have worked on knowledge acquisition and graphical display of process information and on the description and understanding of the results of laboratory experiments. The highlights of this work are summarized in several categories below.

C.1.1 Qualitative Modeling and Simulation

Our work in qualitative simulation has been directed towards building a program which embodies a theory of the tryptophan system. This simulator will contain two different types of qualitative models: a model which inter-relates the values of numeric parameters of the system via constraints, and a model which contains symbolic representations of the structure and function of objects in the system. The implementation of both these models has been partially completed.

The parametric model will be similar to existing qualitative simulation programs (such as those of deKleer, Kuipers, and Iwasaki) in that it focuses upon the relations among a small set of numeric parameters which describe a system. It will be different from these systems in that both its parameters and the relations between them will be more flexible and expressive.

Parameters will be able to take on a different types of values depending on the problem at hand and the degree of precision with which biologists understand a given mechanism, e.g., quantitative (1.23), relative (2 times Normal), or qualitative (High). Relations will be expressed with what we term "malleable constraints". These rectify an important shortcoming we have noted in previous work on qualitative simulation. The motivations behind this work are that qualitative representations should be used instead of quantitative because quantitative information is often unavailable, or can obscure the essence of a problem and make explanation more difficult. Thus, researchers have chosen to let parameters in their models take on qualitative values such as {-1, 0, 1}. We propose that these same concerns apply to the *relations among* parameters as well as their values. Just as we might know the values of different parameters with different degrees of precision, so we might understand the relationship between parameters more or less precisely. Thus we strive to represent constraints themselves in a qualitative fashion when the precise form of a relation is not known, and are studying how to propagate different types of values through these malleable constraints.

We have begun to design a language for expressing constraints with varying degrees of precision. A very imprecise constraint would state only that two parameters vary

directly (or inversely). Adding more precision could tell us the form of the constraint, e.g., is it linear, quadratic, or exponential? More precision still would tell us the value of the constant for a linear constraint.

We have constructed a knowledge base which describes roughly 15 parameters and the relations between them, and we are currently able to propagate qualitative values through this network. Work has begun on representing malleable constraints and other types of values.

The structural model should describe the mechanisms of the trp system at several different levels of detail. This has benefits for both problem solving and discovery.

Processes and objects can be represented at different levels of detail by abstracting away elements of their description. For example, the process of transcription can be thought of as simply consisting of subprocesses of initiation, elongation, and termination, yet each of these subprocesses is known in greater detail. Likewise, the tryptophan operon can be thought of abstractly as a sequence of DNA segments (such as genes), or as a long, detailed list of specific nucleotides. Simpler descriptions can be reasoned about more quickly and will give reliable predictions for certain questions. They also provide an abstraction hierarchy for the discovery process, allowing a discovery module to focus on increasingly detailed sub-parts of a model to efficiently determine where the error in a theory lies.

C.1.2 Process Description and Graphical Display

A system has been built which generalizes our experience in process description by providing a simplified interface for the domain-independent description and animation of process knowledge. The system allows processes to be broken down into component sub-processes and the causal and time-oriented relationships of the subprocesses to be specified. In addition, objects utilized by the processes can be conveniently described and "drawn" with modes and points of interaction among the objects given by the user. All knowledge about processes and objects is automatically stored in the framework of a KEE knowledge base.

After process and object description, the system automatically animates the process by displaying one of several primitive types of interactions among objects in the proper time order dictated by the process knowledge base. This system has been tested on the tryptophan operon domain and its utility is currently being explored in a medical simulation domain.

C.1.3 Acquisition of Experimental Data

The KEE ActiveImages facility has been used to quickly build a convenient interface for biologists to describe the results of genetic regulation experiments. Information is entered by touching (with a mouse-operated cursor) one of several graphical gauges representing such information as absolute and relative cell growth rates. The values provided are automatically checked for consistency and entered in the frame-based experiment description knowledge base. Quantitative values are translated to qualitative values for use by the modeling and simulation system.

C.2 Research in Progress

Over the next several months we will continue our work in qualitative simulation, modeling, and process description. The full qualitative-quantitative parameter model will be completed and tested over a wide range of possible values in the trp operon system. The first version of the imprecise constraint description language will be completed. Further experiments will be performed on reasoning about processes in a hierarchy of abstraction space. The generic process description and animation system will be tested in a different, most likely medical, domain.

D. Publications

1. Bach, R., Friedland, P., Brutlag, D., and Kedes, L.: *MAXIMIZE, a DNA sequencing strategy advisor.* Nucleic Acids Res. 10(1):295-304, January, 1982

2. Bach, R., Friedland, P., and Iwasaki, Y.: *Intelligent computational assistance for experiment design.* Nucleic Acids Res. 12(1):11-29, January, 1984.

3. Brutlag, D., Clayton, J., Friedland, P. and Kedes, L.: *SEQ: A nucleotide sequence analysis and recombination system.* Nucleic Acids Res. 10(1):279-294, January, 1982.

4. Clayton, J. and Kedes, L.: *GEL, a DNA sequencing project management system.* Nucleic Acids Res. 10(1):305-321, January, 1982.

5. Feitelson, J. and Stefik, M.J.: *A case study of the reasoning in a genetics experiment.* Heuristic Programming Project Report HPP-77-18 (working paper), May, 1977.

6. Friedland, P.: *Knowledge-based experiment design in molecular genetics.* Proc. Sixth IJCAI, August, 1979, pp. 285-287.

7. Friedland P.: *Knowledge-based experiment design in molecular genetics.* Stanford Computer Science Report STAN-CS-79-760 (Ph.D. thesis), December, 1979.

8. Friedland, P., Kedes, L. and Brutlag D.: *MOLGEN--Applications of symbolic computation and artificial intelligence to molecular biology.* Proc. Battelle Conference on Genetic Engineering, April, 1981.

9. Friedland, P.: *Acquisition of procedural knowledge from domain experts.* Proc. Seventh IJCAI, August, 1981, pp. 856-861.

10. Friedland, P., Kedes, L., Brutlag, D., Iwasaki, Y. and Bach R.: *GENESIS, a knowledge-based genetic engineering simulation system for representation of genetic data and experiment planning.* Nucleic Acids Res. 10(1):323-340, January, 1982.

11. Friedland, P., and Kedes, L.: *Discovering the secrets of DNA.* Communications of the ACM, 28(11):1164-1186, November, 1985, and IEEE/Computer, 18(11):49:69, November, 1985.

12. Friedland, P. and Iwasaki Y.: *The concept and implementation of skeletal plans.* Journal of Automated Reasoning, 1(2): 161-208, 1985.

13. Friedland, P., Armstrong, P., and Kehler, T.: *The role of computers in biotechnology.* BIO\TECHNOLOGY 565-575, September, 1983.

14. Iwasaki, Y. and Friedland, P.: *SPEX: A second-generation experiment design system.* Proc. of Second National Conference on Artificial Intelligence, August, 1982, pp. 341-344.

15. Martin, N., Friedland, P., King, J. and Stefik M.J.: *Knowledge base management for experiment planning in molecular genetics.* Proc. Fifth IJCAI, August, 1977, pp. 882-887.

16. Meyers, S. and Friedland, P.: *Knowledge-based simulation of regulatory*

genetics in bacteriophage Lambda. Nucleic Acids Res. 12(1):1-9, January, 1984.

17. Stefik, M. and Friedland, P.: *Machine inference for molecular genetics: Methods and applications.* Proc. of NCC, June, 1978.

18. Stefik, M.J. and Martin N.: *A review of knowledge based problem solving as a basis for a genetics experiment designing system.* Stanford Computer Science Report STAN-CS-77-596, March, 1977.

19. Stefik, M.: *Inferring DNA structures from segmentation data: A case study.* Artificial Intelligence 11:85-114, December, 1977.

20. Stefik, M.: *An examination of a frame-structured representation system.* Proc. Sixth IJCAI, August, 1979, pp. 844-852.

21. Stefik, M.: *Planning with constraints.* Stanford Computer Science Report STAN-CS-80-784 (Ph.D. thesis), March, 1980.

E. Funding Support

The MOLGEN grant is titled: MOLGEN: Applications of Artificial Intelligence to Molecular Biology: Research in Theory Formation, Testing, and Modification. It is NSF Grant MCS-8310236. Current Principal Investigators are Edward A. Feigenbaum Professor of Computer Science and Charles Yanofsky, Professor of Biology. MOLGEN is currently funded from 11/85 to 10/86 at $135,000 including indirect costs as the third year of a three year grant.

II. INTERACTIONS WITH THE SUMEX-AIM RESOURCE

SUMEX-AIM continues to provide the bulk of our computing resources. The facility has not only provided excellent support for our programming efforts but has served as a major communication link among members of the project. Systems available on SUMEX-AIM such as INTERLISP, TV-EDIT, and BULLETIN BOARD have made possible the project's programming, documentation and communication efforts. The interactive environment of the facility is especially important in this type of project development.

We strongly approve of the network-oriented approach to a programming environment into which SUMEX has evolved. The ability to utilize LISP workstations for intensive computing while still communicate with all of the other SUMEX resources has been very valuable to our work. We currently have a satisfactory mode of operation where most programming takes place on the workstations and most electronic communications, information sharing, and document preparation takes place within the mature TOPS-20 environment. The evolution of SUMEX has alleviated most of our previous problems with resource loading and file space. Our current workstations are not quite fast nor sophisticated enough, but we are encouraged by the progress that has been made.

We have taken advantage of the collective expertise on medically-oriented knowledge-based systems of the other SUMEX-AIM projects. In addition to especially close ties with other projects at Stanford, we have greatly benefited by interaction with other projects at yearly meetings and through exchange of working papers and ideas over the system.

The ability for instant communication with a large number of experts in this field has been a determining factor in the success of the MOLGEN project. It has made possible the near-instantaneous dissemination of MOLGEN systems to a host of experimental users in laboratories across the country. The wide-ranging input from these users has greatly improved the general utility of our project.

We find it very difficult to find fault with any aspect of the SUMEX resource management. It has made it easy for us to expand our user group, to give demonstrations (through the 20/20 adjunct system as well as the LISP workstations), and to disseminate software to non-SUMEX users overseas.

III. RESEARCH PLANS

A. Project Goals And Plans

Our current work has the following major goals:

1. Use the knowledge base to explain observations that are indeed explainable without changes to the current model. For example, "I have observed a mutation that causes constitutive (uncontrolled) production of tryptophan. How can that be explained within the Jacob-Monod model?" This process will be accomplished by some combination of forward simulation and backward rule-chaining.

2. Begin to recognize when observations are "interesting." Interesting here has one of the following broad meanings:

    a. A seeming direct contradiction to the existing theory.

    b. A statistically rare occurrence (one that is understandable by the current theory, but should not occur very often).

    c. A dramatic confirmation of the existing model.

    d. An observation currently unpredictable by the current model because the model is either not detailed enough or incomplete. The observation in this case must have a relation to the model because an important object of the model is involved or it relates to an effect predicted by the model.

3. Build a mechanism for postulating extensions or corrections to the current theory: a contrained regulatory theory generator. The overall approach to this mechanism is perhaps the most interesting problem in our work. In discussions with other computer scientists, the notion of "or" reasoning where the theory construction process consists of hierarchical refinement of abstract ideas into more detailed ones, and "and" reasoning where the theory is built up in little pieces at many different levels simultaneously has emerged. We see strong evidence for both types of reasoning within Yanofsky's project. In fact, as stated above, the global model of Yanofsky's laboratory is a hybrid one. Individual graduate students performed "and" tasks--filling in details of seemingly unrelated pieces of the model. Yanofsky was the master "or" reasoner, slowly building a hierarchical model of the new regulatory mechanism. It is in this area of our research where the greatest discussion with AI colleagues is needed and which may produce the most significant AI benefits.

4. Build a mechanism for evaluating alternative theories. This would include rating the theories based on plausibility, selectability, completeness, significance, and so on. We hope the evaluation process produces information useful in discriminating among the possible theories.

5. Test the entire structure on the evolving trp operon regulatory system. Experiment with different initial knowledge bases to see how the discovery

process is altered by the availability of new techniques, analogous systems, and so forth.

B. Justification and Requirements for Continued SUMEX Use

The MOLGEN project depends heavily on the SUMEX facility. We have already developed several useful tools on the facility and are continuing research toward applying the methods of artificial intelligence to the field of molecular biology. The community of potential users is growing nearly exponentially as researchers from most of the biomedical-medical fields become interested in the technology of recombinant DNA. We believe the MOLGEN work is already important to this growing community and will continue to be important. The evidence for this is an already large list of pilot exo-MOLGEN users on SUMEX.

We support with great enthusiasm the acquisition of satellite computers for technology transfer and hope that the SUMEX staff continues to develop and support these systems. One of the oft-mentioned problems of artificial intelligence research is exactly the problem of taking prototypical systems and applying them to real problems. SUMEX gives the MOLGEN project a chance to conquer that problem and potentially supply scientific computing resources to a national audience of biomedical-medical research scientists.

Responses to Questions Regarding Resource Future

1. Role of SUMEX after 7/86--I strongly believe that the 2060 should have continuing support for the foreseeable future. The maturity of software for communications, document preparation, and general support of scientific literacy is unsurpassed. One has only to note the heavy continued load on SUMEX, despite the proliferation of workstations, VAXes, etc. around the KSL to see that it is still being used productively. In addition, the ability to easily work from home at all hours contributes greatly to overall productivity within the SUMEX community.

2. Will my group require continued access--Yes, very much so for all of the reasons outlined above.

3. Impact of user fees--Modest user fees would not have an enormous impact, but would prevent the kind of easy, productive use for general purposes that SUMEX now serves. I think the greater impact would be on not fully established or new research groups during start-up mode.

4. Workstation plans--My group, MOLGEN, already makes extensive use of workstations for mainline computing purposes. Despite this use, we still find the SUMEX 2060 invaluable.

I would add to #1, that continuing research on melding together a distributed environment (of which both single-user workstations and the 2060 are parts) should be a major continuing goal of SUMEX research.

IV.A.3. ONCOCIN Project

ONCOCIN Project

Edward H. Shortliffe, M.D., Ph.D.
Departments of Medicine and Computer Science
Stanford University

I. SUMMARY OF RESEARCH PROGRAM

A. Project Rationale

The ONCOCIN Project is one of many Stanford research programs devoted to the development of knowledge-based expert systems for application to medicine and the allied sciences. The central issue in this work has been to develop a program that can provide advice similar in quality to that given by human experts, and to ensure that the system is easy to use and acceptable to physicians. The work seeks to improve the interactive process, both for the developer of a knowledge-based system, and for the intended end user. In addition, we have emphasized clinical implementation of the developing tool so that we can ascertain the effectiveness of the program's interactive capabilities when it is used by physicians who are caring for patients and are uninvolved in the computer-based research activity.

B. Medical Relevance and Collaboration

The lessons learned in building prior production rule systems have allowed us to create a large oncology protocol management system much more rapidly than was the case when we started to build MYCIN. We introduced ONCOCIN for use by Stanford oncologists in May 1981. This would not have been possible without the active collaboration of Stanford oncologists who helped with the construction of the knowledge base and also kept project computer scientists aware of the psychological and logistical issues related to the operation of a busy outpatient clinic.

C. Highlights of Research Progress

C.1 Background and Overview of Accomplishments

The ONCOCIN Project is a large interdisciplinary effort that has involved over 35 individuals since the project's inception in July 1979. The work is currently in its seventh year; we summarize here the milestones that have occurred in the research to date:

- *Year 1:* The project began with two programmers (Carli Scott and Miriam Bischoff), a Clinical Specialist (Dr. Bruce Campbell) and students under the direction of Dr. Shortliffe and Dr. Charlotte Jacobs from the Division of Oncology. During the first year of this research (1979-1980), we developed a prototype of the ONCOCIN consultation system, drawing from programs and capabilities developed for the EMYCIN system-building project. During that year, we also undertook a detailed analysis of the day-to-day activities of the Stanford Oncology Clinic in order to determine how to introduce ONCOCIN with minimal disruption of an operation which is already running smoothly. We also spent much of our time in the first year giving careful consideration to the most appropriate mode of interaction with physicians in order to optimize the chances for ONCOCIN to become a useful and accepted tool in this specialized clinical environment.

- *Year 2:* The following year (1980-1981) we completed the development of a special interface program that responds to commands from a customized keypad. We also encoded the rules for one more chemotherapy protocol (oat cell carcinoma of the lung) and updated the Hodgkin's Disease protocols when new versions of the documents were released late in 1980; these exercises demonstrated the generality and flexibility of the representation scheme we had devised. Software protocols were developed for achieving communication between the interface program and the reasoning program, and we coordinated the printing routines needed to produce hard copy flow sheets, patient summaries, and encounter sheets. Finally, lines were installed in the Stanford Oncology Day Care Center, and, beginning in May 1981, eight fellows in oncology began using the system three mornings per week for management of their patients enrolled in lymphoma chemotherapy protocols.

- *Year 3:* During our third year (1981 - 1982) the results of our early experience with physician users guided both our basic and applied work. We designed and began to collect data for three formal studies to evaluate the impact of ONCOCIN in the clinic. This latter task required special software development to generate special flow sheets and to maintain the records needed for the data analysis. Towards the end of 1982 we also began new research into a *critiquing model* for ONCOCIN that involves "hypothesis assessment" rather than formal advice giving. Finally, in 1982 we began to develop a query system to allow system builders as well as end users to examine the growing complex knowledge base of the program.

- *Year 4:* Our fourth year (1982-1983) saw the departure of Carli Scott, a key figure in the initial design and implementation of ONCOCIN, the promotion of Miriam Bischoff to Chief Programmer, and the arrival of Christopher Lane as our second scientific programmer. At this time we began exploring the possibility of running ONCOCIN on a single-user professional workstation and experimented with different options for data-entry using a "mouse" pointing device. Christopher Lane became an expert on the Xerox workstations that we are using. In addition, since ONCOCIN had grown to such a large program with many different facets, we spent much of our fourth year documenting the system. During that year we also modified the clinic system based upon feedback from the physician-users, made some modifications to the rules for Hodgkin's disease based upon changes to the protocols, and completed several evaluation studies.

- *Year 5:* The project's fifth year (1983-1984) was characterized by growth in the size of our staff (three new full-time staff members and a new oncologist joined the group). The increased size resulted from a DRR grant that permitted us to begin a major effort to rewrite ONCOCIN to run on professional workstations. Dr. Robert Carlson, who had been our Clinical Specialist for the previous two years, was replaced by Dr. Joel Bernstein, while Dr. Carlson assumed a position with the nearby Northern California Oncology Group; this appointment permitted him to continue his affiliation both with Stanford and with our research group. In August of 1983, Larry Fagan joined the project to take over the duties of the ONCOCIN Project Director while also becoming the Co-Director of the newly formed Medical Information Sciences Program. Dr. Fagan continues to be in charge of the day-to-day efforts of our research. An additional programmer, Jay Ferguson, joined the group in the fall to assist with the effort required to transfer ONCOCIN from SUMEX to the 1108 workstation. A fourth programmer, Joan Differding, joined the staff to work on our protocol acquisition effort (OPAL).

- *Year 6:* During our sixth year (1984-1985) we have further increased the size of our programming staff to help in the major workstation conversion effort. The ONCOCIN and OPAL efforts were greatly facilitated by a successful application for an equipment grant from Xerox Corporation. With a total of 15 Xerox LISP machines now available for our group's research, all full time programmers have dedicated machines, as do several of the senior graduate students working on the project. Christopher Lane took on full-time responsibility for the integration and maintenance of the group's equipment and associated software. Two of our programming staff moved on to jobs in industry (Bischoff and Ferguson) and three new programmers (David Combs, Cliff Wulfman, and Samson Tu) were hired to fill the void created by their departure and by the reassignment of Christopher Lane.

  In addition to funding from DRR for the workstation conversion effort, we have support from the National Library of Medicine which supports our more basic research activities regarding biomedical knowledge representation, knowledge acquisition, therapy planning, and explanation as it relates to the ONCOCIN task domain. We have continued to study the therapy planning process under support from the NLM. This research is led by Dr. Fagan and has concentrated on how to represent the therapy-planning strategies used to decide treatment for patients who run into serious problems while on protocol-described treatment. The physicians who treat these patients often seek out a consultation with the protocol study chairman. Dr. Branimar Sikic, a faculty member from the Stanford University Department of Medicine, and the Study Chairman for the oat cell protocol, is collaborating on this project. Janice Rohn joined the ONCOCIN project as data manager and to assist in the knowledge entry process.

- *Year 7:* This year (1985-86) marked several milestones in our research on workstation-based programming. The OPAL knowledge acquisition system became operational, and several new oncology protocols have been entered using this system. David Combs has been primarily responsible for creating the operational version of OPAL (based on the initial prototype by Joan Differding Walton). As anticipated, we have increased the speed and ease with which protocols can be added to the ONCOCIN knowledge base.

  Based on the protocols entered through OPAL, we have begun experimental testing of the workstation version of ONCOCIN in the Stanford oncology clinic. Clifford Wulfman developed the user interface (based on an initial prototype designed by Christopher Lane). Samson Tu developed the reasoning component (designed originally by Jay Ferguson). Much of their work is built upon an object-oriented system developed for our group by Christopher Lane. We have connected the various parts of the system, and have demonstrated that we have the capability to run ONCOCIN with the reasoning program and interface program on different machines in the communication network. The current version of the program is currently run on a single workstation, but future versions may take advantage of the multiple machine option. To increase the speed at which we are able to test protocols entered into ONCOCIN, we have developed additional programs to test real and synthetic cases without user interaction; these are then reviewed by our collaborating clinicians.

  We have also developed a workstation-based program, OPUS, to help clinicians determine which protocols are appropriate for specific patients. OPUS was designed and implemented by Janice Rohn with the assistance of

Christopher Lane. We have been using it in the clinic setting since the end of 1985. Thus, in addition to providing an information resource about protocols, the use of a graphically-oriented program provided a way to learn about the software style and hardware used in the workstation version of ONCOCIN.

We discontinued the mainframe version of ONCOCIN, and are using the workstation version exclusively. The performance of the mainframe version of ONCOCIN was documented in two evaluation papers that appeared in clinical journals (see Hickam and Kent papers).

We have continued our basic research in the design of advanced therapy-planning programs: the ONYX project. We have developed a model for planning which includes techniques from the fields of artificial intelligence, simulation, and decision analysis. Artificial intelligence techniques are used to create a small number of possible plans given the ideal therapy and the patient's past treatment history. Simulation techniques and decision analysis are used to examine and order the most promising plans. Our goal is to allow ONCOCIN to give advice in a wider range of situations; in particular, the system should be able to recommend plans for patients who have an unusual response to chemotherapy.

During this year, Stephen Rappaport, M.D. joined us as a programmer on the therapy planning research. Clinical expertise for ONCOCIN was provided by Richard Lenon, M.D. and Robert Carlson, M.D.

C.2 Research in Progress

Our research in the ONCOCIN project over the last year comprised three major categories: (1) conversion of ONCOCIN to the workstation version, (2) development of a knowledge acquisition interface (OPAL) for entering new protocols, and (3) modeling of the strategic therapy selection process (ONYX). We are now able to explore ways to test the system beyond the Stanford environment.

A summary of our current research endeavors follows.

C.2.1 Transfer of the ONCOCIN system from the DEC-20 to the Xerox 1100 Series machines

During the process of converting to the workstation version of ONCOCIN, we redesigned segments of the program. We have completed the major portion of that work, and our experience with the new version has suggested additional areas for improving the reasoning techniques and knowledge representation of ONCOCIN.

- *Redesign of the reasoning component.* A major impetus for the redesign of the system was to develop more efficient methods to search the knowledge base during the running of a case. We have implemented a reasoning program that uses a discrimination network to process the cancer protocols. This network provides for a compact representation of information which is common to many protocols but does not require the program to consider and then disregard information related to protocols that are irrelevant to a particular patient. We continue to improve portions of the reasoning component that are associated with reasoning over time; e.g., modeling the appropriate timing for ordering tests and identifying the information which needs to be gathered before the next clinic visit. In general, we are concentrating on improving the representation of the knowledge regarding sequences of therapy actions specified by the protocol.

We are also improving the reasoning component's efficiency, such as by building concurrency into the reasoning process. Our overall aim in this area is to increase generality of the reasoning program; our long-term goal is to develop E-ONCOCIN, a domain-independent, time-oriented reasoning program.

- *Development of a temporal network.* The ability to represent temporal information is a key element of programs that must reason about treatment protocols. The earlier version of the ONCOCIN system did not have an explicit structure for reasoning about time-oriented events. We are experimenting with different configurations of the temporal network, and with the syntax for querying the network. We are also adapting this network so that it can interface with the ONYX therapy-planning systems.

- *Extensions to the user interface.* We continue to experiment with various configurations of the user interface, including new formats for reviewing the therapy recommendations and specifying tests to be ordered. We have also developed a program which allows the data manager and end users to start up their interaction with ONCOCIN.

  A continuing area of research concerns how to guide the user to the most appropriate items to enter (based on the needs of the reasoning program) without disrupting the fixed layout of the flowsheet. The mainframe version of ONCOCIN modified the flowsheet in order to extract necessary information from the user. In the workstation version, we have developed a guidance mechanism which alerts the user to items that are needed by the reasoning program. The user is not required to deviate from a preferred order of entry nor required to respond to a question for which no current answer is available.

- *System support for the reorganization.* The LISP language, which we used to build the first version of ONCOCIN, does not explicitly support basic knowledge manipulation techniques (such as message passing, inheritance techniques, or other object-oriented programming structures). These facilities are available in some commercial products, but none of the existing commercial implementations provide the reliability, speed, size, or special memory-manipulation techniques that are needed for our project. We have therefore developed a "minimal" object-oriented system to meet our specifications. The object system is currently in use by each component of the new version of ONCOCIN and in the software used to connect these components. In addition, all ONCOCIN student projects are now based on this programming environment.

C.2.2 Interactive Entry of Chemotherapy Protocols by Oncologists (OPAL)

A major effort in this grant year has been the development and testing of software (the OPAL system) that will permit physicians who are not computer programmers to enter protocol information on a structured set of forms presented on a graphics display. Most expert systems require tedious entry of the system's knowledge. In many other medical expert systems, each segment of knowledge is transferred from the physician to the programmer, who then enters the knowledge into the expert system. We have taken advantage of the generally well-structured nature of cancer treatment plans to design a knowledge entry program that can be used directly by clinicians. The structure of cancer treatment plans includes:

- choosing among multiple protocols (that may be related to each other);

- describing experimental research arms in each protocol;
- specifying individual drugs and drug combinations;
- setting the drug dosage level;
- and modifying either the choice of drugs or their dosage.

Using the graphics-oriented workstations, this information is presented to the user as computer-generated forms which appear on the screen. After the user fills in the blanks on the forms, the program generates the rules used to drive the reasoning process. As the user describes more detailed aspects of the protocol, new forms are added to the computer display; these allow the user to specify the special cases that make the protocols so complicated. Although the user is unaware of the creation of the knowledge base from the interaction with OPAL, a complex set of translations are taking place. The user's entries are mapped into an intermediate data structure (IDS) that is common for all protocols. From the IDS, a translation program generates rules for creating and modifying treatment, and integrates them with the existing ONCOCIN knowledge base. Improving the design of the IDS and the rule translation programs will be a major research effort of this year.

Although the "forms" were specifically designed for cancer treatment plans, the techniques used to organize data can be extended to other clinical trials, and eventually to other structured decision tasks. The key factor is to exploit the regularities in the structure of the task (e.g., this interface has an extensive notion of how chemotherapy regimens are constructed) rather than to try to build a knowledge-entry system that can accept *any* possible problem specification. The OPAL program is based upon a domain-independent forms creation package designed and implemented by David Combs. This program will provide the basis for our extension of OPAL to other application areas.

We have entered six protocols covering many different organ systems and styles of protocol design. Based on this experience, we are modifying OPAL to increase the percentage of the protocol that can be entered directly by our clinical collaborators. One direction in which we have extended the OPAL program is in providing a graphical interface of nodes and arcs to specify the procedural knowledge about the order of treatments and important decision points within the treatments. This work is described in several papers by Musen.

C.2.3 Strategic Therapy Planning (ONYX)

As mentioned above, we have continued our research project (ONYX) to study the therapy-planning process and to determine how clinical strategies are used to plan therapy in unusual situations. Our goals for ONYX are: (1) to conduct basic research into the possible representations of the therapy-planning process, (2) to develop a computer program to represent this process, and (3) eventually to interface the planning program with ONCOCIN. The project members (Fagan, Kahn, Langlotz, Rappaport, and Tu) have spent many hours meeting with Dr. Sikic to determine how he plans therapy for patients whose special clinical situation precludes following the standard therapeutic plan described in the protocol document.

The prototype program design has four components: (1) to review the patient's past record and recognize emerging problems, (2) to formulate a small number of revised therapy plans based on existing problems, (3) to determine the results of the generated plans by using simulation, and (4) to weight the results of the simulation and rank

order the plans by performing decision analysis. This model is described in the papers by Langlotz.

C.2.4 Documentation

We recently videotaped a lecture and demonstration of the ONCOCIN and OPAL systems at the XEROX Palo Alto Research Center. This videotape will be available for loan from our offices. Our previous videotapes have been shown at scientific meetings and have been distributed to many researchers in other countries. The publications described below further document our recent work on ONCOCIN.

C.2.5 Dissemination

We are planning experimental installation of ONCOCIN workstations in private oncology offices in San Jose and San Francisco. An application proposing this project is currently under review.

D. Publications Since January, 1985

1. Hickam, D.H., Shortliffe, E.H., Bischoff, M.B., Scott, A.C., Jacobs, C.D. A study of the treatment advice of a computer-based cancer chemotherapy protocol advisor (Memo KSL-85-21). Annals of Internal Medicine 103(6 pt 1):928-936 (1985).

2. Kent, D.L., Shortliffe, E.H., Carlson, R.W., Bischoff, M.B., Jacobs, C.D. Improvements in data collection through physician use of a computer-based chemotherapy treatment consultant (Memo KSL-85-22). Journal of Clinical Oncology 3:1409-1417 (1985).

3. Tsuji, S. and Shortliffe, E.H. Graphical access to a medical expert system: I. Design of a knowledge engineer's interface (Memo KSL-85-11). Meth. Inf. Med., Vol. 25, April 1986.

4. Preston, K.,Jr., Fagan, L.M., Huang, H.K., Pryor, T.A. Computing in medicine. IEEE Computer 17(10):294-313, October 1984.

5. Langlotz, C., Fagan, L., Tu, S., Williams, J., Sikic, B. ONYX: An architecture for planning in uncertain environments (Memo KSL-85-10). Proceedings of the Ninth International Joint Conference on Artificial Intelligence, pp.447-449, Los Angeles, CA, August 1985.

6. Lane, C.D., Differding, J.C., Shortliffe, E.H. Graphical access to a medical expert system: II. Design of an interface for physicians (Memo KSL-85-15). To appear in Meth. Inf. Med., July 1986.

7. Musen, M., Langlotz, C., Fagan, L., Shortliffe, E.H. Rationale for knowledge base redesign in a medical advice system (Memo KSL-85-17). Proceedings of AAMSI Congress 85 (A. Levy and B. Williams, Eds.), pp. 197-201, San Francisco, May 20-22, 1985.

8. Fagan, L. New directions for expert systems: Examples from the ONCOCIN project. Proceedings of AAMSI Congress 85 (A. Levy and B. Williams, Eds.), pp. 183-186, San Francisco, May 20-22, 1985.

9. Kahn, M.G., Ferguson, J. C., Shortliffe, E.H., Fagan, L. Representation and Use of Temporal Information in ONCOCIN. (Memo KSL-85-8). Proceedings of the 9th SCAMC, pp. 172-176, Baltimore, MD, November 1985.

10. Fagan, L., Differding, J., Langlotz, C., Tu, S. Knowledge acquisition and strategic therapy planning for cancer clinical trials. Proceedings of the International Conference on Artificial Intelligence in Medicine (I. De Lotto and M. Stefanelli, Eds.), Pavia, Italy, 13-14 September 1985.

11. Musen, M.A., Rohn, J.A., Fagan, L.M., Shortliffe, E.H. Knowledge engineering for a clinical trial advice system: uncovering errors in protocol specification (Memo KSL-85-51). Proceedings of AAMSI Congress 86, Anaheim, CA, May 8-10, 1986.

12. Langlotz, C.P., Fagan, L.M., Shortliffe, E.H. Overcoming limitations of artificial intelligence planning techniques (Memo KSL-85-52). Proceedings of AAMSI Congress 86, Anaheim, CA, May 8-10, 1986.

13. Musen, M.A., Fagan, L.M., Shortliffe, E.H. Graphical specification of procedural knowledge for an expert system (Memo KSL-85-53). December 1985. To be presented at the Second IEEE Computer Society Workshop on Visual Languages, Dallas, TX, June 1986.

14. Combs,D.M., Musen, M.A., Fagan, L.M., Shortliffe, E.H. Graphical entry of procedural and inferential knowledge (Memo KSL-85-56). Proceedings of AAMSI Congress 86, Anaheim, CA, May 8-10, 1986.

15. Lane, C.D., Frisse, M.E., Fagan, L.M., and Shortliffe, E.H. Object-oriented graphics in medical interface design (Memo KSL-85-58). Proceedings of AAMSI Congress 86, Anaheim, CA, May 8-10, 1986.

16. Musen, M.A., Fagan, L.M., Combs, D.M., Shortliffe, E.H. Facilitating knowledge entry for an oncology therapy advisor using a model of the application area (Memo KSL-86-1). To appear in Proceedings of MEDINFO-86, October 1986.

17. Langlotz, C.P., Fagan, L.M., Tu, S.W., Sikic, B.I., Shortliffe, E.H. Combining artificial intelligence and decision analysis for automated therapy planning assistance (Memo KSL-86-3). To appear in Proceedings of MEDINFO-86, October 1986.

18. Kahn, M.G., Fagan, L.M., Shortliffe, E.H. Context-specific interpretation of patient records for a therapy advice system (Memo KSL-86-4). To appear in Proceedings of MEDINFO-86, October 1986.

19. Langlotz, C.P., Shortliffe, E.H., Fagan, L.M. Using decision theory to justify heuristics (Memo KSL-86-26). To appear in Proceedings of AAAI-86, August 1986.

20. Shortliffe, E.H. Artificial Intelligence in Management Decisions: ONCOCIN (Memo KSL-86-39). Proceedings of a Conference on Medical Information Sciences, University of Texas Health Sciences Center at San Antonio, July 1985.

E. Funding Support

Grant Title: "Studies in the Dissemination of Consultation Systems"
Principal Investigator: Edward H. Shortliffe

Agency: Biotechnology Resources Program, Division of Research Resources
ID Number: RR 01613
Term: July 1983 to June 1986
Total award: $624,455
Current award: (7/85-6/86): $200,302

Grant Title: "Therapy-planning strategies for consultation by computer"
Principal Investigator: Edward H. Shortliffe
Agency: National Library of Medicine
ID Number: LM-04136
Term: August 1983 to July 1986
Total award: $211,851
Current award: (8/85-7/86) $74,150

Grant Title: "Knowledge Management for Clinical Trial Advice Systems"
Principal Investigator: Edward H. Shortliffe
Agency: National Library of Medicine
ID Number: 1 R01 LM04420-01
Term: September 1985 through August 1988
Total award: $314,707
Current Award: (9/85-8/86): $95,205

Grant Title:"Information Structure and Use in Knowledge-based Expert Systems"
Principal Investigator: Bruce G. Buchanan
Co-Principal Investigator: Edward H. Shortliffe
Agency: National Science Foundation
ID Number: IST 83-12148
Term: March 1, 1984 - February 28, 1987
Total award: $330,000 (includes indirects)
Current award (3/85-2/86) $52,679 (Shortliffe portion)

Grant Title: Postdoctoral Training in Medical Information Science
Principal Investigator: Edward H. Shortliffe
Agency: National Library of Medicine
ID Number: 1 T32 LM07033
Term: July 1, 1984 - June 30, 1989
Total award: $903,718
Current award (7/1/85-6/30/86): $215,850

Grant Title: Henry J. Kaiser Faculty Scholar in General Internal Medicine
Principal Investigator: Edward H. Shortliffe
Agency: Henry J. Kaiser Family Foundation
Term: July 1983 to June 1988
Total award: $250,000 ($50,000 annually).

Grant Title: "Explanation of Computer-Assisted Therapy Plans"
Principal Investigator: Lawrence M. Fagan
Agency: NIH
ID Number: LM04316
Term: Feb. 1985 to Jan. 1988
Total award: $107,441.
Current award: (2/85-1/86) $37,500

II. INTERACTIONS WITH THE SUMEX-AIM RESOURCE

A. Medical Collaborations and Program Dissemination via SUMEX

A great deal of interest in ONCOCIN has been shown by the medical, computer science, and lay communities. We are frequently asked to demonstrate the program to Stanford visitors (both the prototype system running in the clinic and the newer work transferring the system to professional workstations). We also demonstrated our developing workstation code in the Xerox exhibit in the trade show associated with AAAI-84 in Austin, Texas and IJCAI-85 in Los Angeles. Physicians have generally been enthusiastic about ONCOCIN's potential. The interest of the lay community is reflected in the frequent requests for magazine interviews and television coverage of the work. Articles about MYCIN and ONCOCIN have appeared in such diverse publications as *Time* and *Fortune*, and ONCOCIN has been featured on the "NBC Nightly News," the PBS "Health Notes" series, and "The MacNeil-Lehrer Report." Due to the frequent requests for ONCOCIN demonstrations, we have produced a videotape about the ONCOCIN research which includes demonstrations of our professional workstation research projects and the 2020-based clinic system. The tape has been shown at several national meetings, including the 1984 Workshop on Artificial Intelligence in Medicine, the 1984 meeting of the Society for Medical Decision Making, and the 1985 meeting of the Society for Research and Education in Primary Care Internal Medicine. The tape has also been shown to both national and international researchers in biomedical computing. We have also completed an updated tape of our activities for demonstration purposes.

Our group also continues to oversee the MYCIN program (not an active research project since 1978) and the EMYCIN program. Both systems continue to be in demand as demonstrations of expert systems technology. MYCIN has been demonstrated via networks at both national and international meetings in the past, and several medical school and computer science teachers continue to use the program in their computer science or medical computing courses. Researchers who visit our laboratory often begin their introduction by experimenting with the MYCIN/EMYCIN systems. We also have made the MYCIN program available to researchers around the world who access SUMEX using the GUEST account. EMYCIN has been made available to interested researchers developing expert systems who access SUMEX via the CONSULT account. One such consultation system for psychopharmacological treatment of depression, called Blue-Box (developed by two French medical students, Benoit Mulsant and David Servan-Schreiber), was reported in July of 1983 in *Computers and Biomedical Research.*

B. Sharing and Interaction with Other SUMEX-AIM Projects

The community created on the SUMEX resource has other benefits which go beyond actual shared computing. Because we are able to experiment with other developing systems, such as INTERNIST/CADUCEUS, and because we frequently interact with other workers (at AIM Workshops or at other meetings), many of us have found the scientific exchange and stimulation to be heightened. Several of us have visited workers at other sites, sometimes for extended periods, in order to pursue further issues which have arisen through SUMEX- or workshop-based interactions. In this regard, the ability to exchange messages with other workers, both on SUMEX and at other sites, has been crucial to rapid and efficient dissemination of ideas. Certainly it is unusual for a small community of researchers with similar scholarly interests to have at their disposal such powerful and efficient communication mechanisms, even among those researchers on opposite coasts of the country.

During this past year, we have had extensive interactions with Randy Miller at

Pittsburgh. Via floppy disks and SUMEX, we have experimented with several versions of the QUIK program. The interaction was very much facilitated by the availability of SUMEX for communication and data transmission.

C. Critique of Resource Management

Our community of researchers has been extremely fortunate to work on a facility that has continued to maintain the high standards that we have praised in the past. The staff members are always helpful and friendly, and work as diligently to please the SUMEX community as to please themselves. As a result, the computer is as accessible and easy-to-use as they can make it. More importantly, it is a reliable and convenient research tool. We extend special thanks to Tom Rindfleisch for maintaining such high professional standards. As our computing needs grow, we have increased our dependence on special SUMEX skills such as networking and communication protocols.

III. RESEARCH PLANS

A. Project Goals and Plans

In the coming year, there are several areas in which we expect to expend our efforts on the ONCOCIN System:

1. *Development of a workstation model for cost-effective dissemination of clinical consultation systems.* To meet this specific aim we will continue the basic and applied programming efforts (ONCOCIN, OPAL, and ONYX) described earlier in this report.

2. *To encode and implement for use by ONCOCIN the commonly used chemotherapy protocols from our oncology clinic.* In the upcoming year, we will:

    - Extend the OPAL protocol entry system
    - Continue entry of additional protocols, hopefully at the rate of one protocol/month (including testing)

3. *To continue testing of the workstation version of ONCOCIN.*

4. *To generalize the reasoning and interaction components of the ONCOCIN system for other applications.*

B. Justification and Requirements for Continued SUMEX Use

All the work we are doing (ONCOCIN plus continued use of the original MYCIN program) continues to be dependent on daily use of the SUMEX resource. Although much of the ONCOCIN work is shifting to Xerox workstations, the SUMEX 2060 and the 2020 continue to be key elements in our research plan. The programs all make assumptions regarding the computing environment in which they operate.

In addition, we have long appreciated the benefits of GUEST and network access to the programs we are developing. SUMEX greatly enhances our ability to obtain feedback from interested physicians and computer scientists around the country. Network access has also permitted high quality formal demonstrations of our work both from around the United States and from sites abroad (e.g., Finland, Japan, Sweden, Switzerland).

The main development of our project will continue to take place on LISP machines which we have purchased or which have been donated by the XEROX Corporation.

C. Requirements for Additional Computing Resources

The acquisition of the DEC 2020 by SUMEX was crucial to the growth of our research work. It has insured high quality demonstrations and has enabled us to develop a system (ONCOCIN) for real-world use in a clinical setting. As we have begun to develop systems that are potentially useful as stand-alone packages (i.e., an exportable ONCOCIN), the addition of personal workstations has provided particularly valuable new resources. We have made a commitment to the smaller Interlisp-D machines ("D-machines") produced by Xerox, and our work will increasingly transfer to them over the next several years. Our current funding supports our effort to implement ONCOCIN on workstations in the Stanford oncology clinic (and eventually to move the program to non-Stanford environments), but we will simultaneously continue to require access to Interlisp on upgraded workstations for extremely CPU-intensive tasks. Although our dependence on SUMEX for workstations has decreased due to a recent gift from XEROX, our requirements for network support of the machines has drastically increased. Individual machines do not provide sufficient space to store all of the software used in our project, nor to provide backup or long-term storage of work in progress. It is the networks, file storage devices, protocol converters, and other parts of the SUMEX network that hold our project together. In addition, with a research group of about 20 people, we are taking advantage of file sharing, electronic mail, and other information coordinating activities provided by the DEC 2060. We hope that with systems support and research by SUMEX staff, we will be able to gradually move away from a need for the central coordinating machine over the next five years.

The acquisition of the DEC 2060, coupled with our increasing use of workstations, has greatly helped with the problems in SUMEX response time that we had described in previous annual reports. We are extremely grateful for access both to the central machine and to the research workstations on which we are currently building the new ONCOCIN prototype. The D-machine's greater address space is permitting development of the large knowledge base that ONCOCIN requires. The graphics capability of the workstations has also enabled us to develop new methods for presenting material to naive users. In addition, the workstations have provided a reliable, constant "load-average" machine for running experiments with physicians and for development work. The development of ONCOCIN on the D-machine will demonstrate the feasibility of running intelligent consultation systems on small, affordable machines in physicians' offices and other remote sites.

D. Recommendations for Future Community and Resource Development

SUMEX is providing an excellent research environment and we are delighted with the help that SUMEX staff have provided implementing enhanced system features on the 2060 and on the workstations. We feel that we have a highly acceptable research environment in which to undertake our work. Workstation availability is becoming increasingly crucial to our research, and we have found over the past year that workstation access is at a premium. The SUMEX staff has been very helpful and understanding about our needs for workstation access, allowing us D-machine use wherever possible, and providing us with systems-level support when needed. We look forward to the arrival of additional advanced workstations and the development of a more distributed computing environment through SUMEX-AIM.

Responses to Questions Regarding Resource Future

"What do you think the role of the SUMEX-AIM resource should be for the period after 7/86, e.g., continue like it is, discontinue support of the central machine, act as a communications crossroads, develop software for user community workstations, etc.?"

We believe that the trend towards distributed computing that characterized the early

1980's will continue during the second half of the decade. Although we have begun this process by moving much of our research activity to LISP machines, the SUMEX DEC-20 continues to be a major source of support for all communication, collaboration, and administrative functions. It also continues to provide a quality LISP environment for rapid prototyping, student projects in the early stages before workstations are made available, and for demonstrating system features to people at a distance. These latter functions are still not well handled by distributed machines, and we believe that a logical role for the resource in the future is to develop software and communications techniques that will allow us to further decrease our dependence on the large central machine.

"Will you require continued access to the SUMEX-AIM 2060 and if so, for how long?"

As indicated above, our needs could still be met with a gradual phaseout of the 2060 over the next 3-5 years, provided that current services such as file handling and backup, mail, document preparation, and advanced network support are available from other machines (e.g., SAFE file server plus the Medical Computer Science file server). This implies maintenance of an ARPANET connection, connections to other campus machines, and facilities for linking together the heterogeneous collection of computing equipment upon which our research group depends. SUMEX would need to concentrate on providing software support for networks and systems software for workstations if it were to provide the same level of service we now experience while moving to a fully distributed environment.

"What would be the effect of imposing fees for using SUMEX resources (computing and communications) if NIH were to require this?"

Since all our research is NIH-supported, we see nothing but administrative headaches without benefits if there were to be a move to require fee-for-service billing for access to shared SUMEX resources. The net effect would simply be a transfer of funds from one arm of NIH to another (assuming that the agencies that currently fund our work could supplement our grants to cover SUMEX charges), and there would be a simultaneous restraining effect on the research environment. The current scheme permits experimentation and flexibility in use that would be severely inhibited if all access incurred an incremental charge.

"Do you have plans to move your work to another machine workstation and if so, when and to what kind of system?"

As mentioned above, and described in greater detail in our annual report, we are making a major effort to move much of our research activity to LISP machines (currently Xerox 1108's, 1186's and HP-9836's). Our familiarity with this technology, and our commitment to it, have resulted solely from the foresight of the SUMEX resource in anticipating the technology and providing for it at the time of their last renewal. However, for the reasons mentioned above, we continue to depend upon the central communication node for many aspects of our activities and could effectively adapt to its demise only if the phaseout were gradual and accompanied by improved support for a totally distributed computing environment.

IV.A.4. PROTEAN Project

PROTEAN Project

Oleg Jardetzky
Nuclear Magnetic Resonance Lab, School of Medicine
Stanford University

Bruce Buchanan, Ph.D.
Computer Science Department
Stanford University

I. SUMMARY OF RESEARCH PROGRAM

A. Project Rationale

The goals of this project are related both to biochemistry and artificial intelligence: (a) use existing AI methods to aid in the determination of the 3-dimensional structure of proteins in solution (not from x-ray crystallography proteins), and (b) use protein structure determination as a test problem for experiments with the AI problem solving structure known as the Blackboard Model. Empirical data from nuclear magnetic resonance (NMR) and other sources may provide enough constraints on structural descriptions to allow protein chemists to bypass the laborious methods of crystallizing a protein and using X-ray crystallography to determine its structure. This problem exhibits considerable complexity. Yet there is reason to believe that AI programs can be written that reason much as experts do to resolve these difficulties [8].

B. Medical Relevance

The molecular structure of proteins is essential for understanding many problems of medicine at the molecular level, such as the mechanisms of drug action. Using NMR data from proteins in solution will allow the study of proteins whose structure cannot be determined with other techniques, and will decrease the time needed for the determination.

C. Highlights of Progress

We have constructed a prototype of such a program, called PROTEAN, designed on the blackboard model [3] [4]. It is implemented in BB1 [5], a framework system for building blackboard systems that control their own problem-solving behavior [6] (see discussion of BB1 above). The reasoning component of PROTEAN directs the actions of the Geometry System (GS) [1], a set of programs that performs the computationally intensive task of positioning portions of a molecule with respect to each other in three dimensions. Currently we have implemented two versions of the GS: an InterLISP version used for quickly testing ideas and developing prototypes of geometric routines on a LISP workstation; and a high performance version written in C and running in the UNIX environment, providing efficient computations on a VAX 11/780. We have coupled the reasoning and geometry programs with an IRIS graphics terminal (shared with SUMEX) which displays the evolving protein structures at several levels of detail. The display in three dimensions is essential to understanding the behavior of the reasoning and geometry systems, and provides essential insights on the problem solving process.

PROTEAN embodies the following experimental techniques for coping with the complexities of constraint satisfaction:

1. The problem-solver partitions each problem into a network of loosely-coupled sub-problems. PROTEAN partitions the problem of positioning all of a protein's constituent structures within a global coordinate system into sub-problems of positioning individual pieces of structures and their immediate neighbors within local coordinate systems. It subsequently composes the most constrained partial solutions developed for these sub-problems in a complete solution for the entire protein. This partitioning and composition technique reduces the combinatorics of search. It also introduces additional constraints in the global characteristics of internally constrained partial solutions. For example, the conformations of partial protein solutions constrain their composability with other partial solutions. In addition, constraints on the overall dimensions of the protein from scattering experiments and indications of which atoms are on the surface of the molecule are used to further limit the possible structures.

2. The problem-solver attempts to solve sub-problems and coordinate solutions at multiple levels of abstraction, where lower levels of abstraction partition solution elements with finer granularity. For example, PROTEAN operates at three levels of abstraction. At the "Solid" level, it positions elements of the protein's secondary structure: alpha-helices, beta-sheets, and random coils. At the "Superatom" level, it positions elements of the protein's primary structure of amino acids: peptide units and side-chains. At the "Atom" level, it positions the protein's individual atoms. Partial solutions at higher levels of abstraction reduce the combinatorics of search at lower levels. Conversely, tightly constrained partial solutions at lower levels introduce new constraints on higher-level solutions.

3. The problem-solver forbears hypothesizing specific partial solutions for a sub-problem in favor of preserving the "family" of solutions consistent with all constraints applied thus far. For example, in positioning a helix within a partial solution, PROTEAN does not attempt to identify a unique spatial position for the helix. Instead, it identifies the entire spatial volume within which the helix might lie, given the constraints applied thus far. Preserving the family of legal solutions accommodates problems with incomplete constraints; the solution is only as constrained as the data are constraining. It also accommodates incompatible constraints by permitting disjunctive sub-families. For PROTEAN, disjunctive sub-volumes imply that the associated structure lies within any one of the sub-volumes or, if the structure is mobile, that it may move from one sub-volume to another.

4. The problem-solver applies constraints one at a time, successively restricting the family of solutions hypothesized for different sub-problems. PROTEAN successively applies constraints on the positions of protein structures, successively restricting the spatial volumes within which they may lie. Independent application of different constraints finesses the problem of integrating qualitatively different kinds of constraints by simply integrating their results. In addition, successive restriction of the family of solutions obviates guessing which specific solutions within a family are likely to be consistent with subsequently applied constraints and the otherwise inevitable back-tracking.

5. The problem-solver tolerates overlapping solutions for different sub-problems. For example, in identifying the volume within which structure-a might lie in partial solution 1, PROTEAN may include part of the volume identified for structure-b. Toleration of overlapping partial solutions is another accommodation of incomplete or incompatible constraints and

potentially dynamic solutions. For PROTEAN, overlapping volumes for two protein structures indicate either: (a) that the two structures actually occupy disjoint sub-volumes that cannot be distinguished within the larger, overlapping volumes identified for them because the constraints are incomplete; or (b) that the two structures are mobile and alternately occupy the shared volume.

6. The problem-solver reasons explicitly about control of its own problem-solving actions: which sub-problems it will attack, which partial solutions it will expand, and which constraints it will apply. Control reasoning guides the problem-solver to perform actions that minimize computation, while maximizing progress toward a complete solution (see section 3.2.1). It also provides a foundation for the problem-solver's explanation of problem-solving activities and intermediate partial solutions (see section 3.2.2) and for its learning of new control heuristics (see section 5.5).

The current version of PROTEAN has five domain knowledge sources that demonstrate the reasoning techniques described above for the assembly of a protein. Each domain knowledge source directs a small portion of the construction of the molecule. These knowledge sources develop partial solutions that position alpha helices, beta strands, and random coils at the Solid level and refine the resulting state families using all available distance constraints.

PROTEAN now uses five control knowledge sources to guide the assembly of a solution to a protein structure. These knowledge sources determine which of the possible assembly actions is the best to perform at each stage of the problem solving.

Proposed work will introduce knowledge sources that apply the reasoning techniques for surface and volume constraints, as well as the ability to reason at Superatom and Atom levels. We also will investigate emergent constraints entailed in reliable partial solutions, composition of partial solutions into complete solutions, and intelligent control.

Multiple blackboards in PROTEAN allow several sets of knowledge to be used. A biochemical knowledge base stores information about proteins and secondary structures, amino acids, and atoms. The problem blackboard describes the protein to be solved and all experimental data observed for the molecule. The evolving solution of the protein structure is built on a third solution blackboard.

The PROTEAN system [2] has been used to construct a complete solution at the solid level of detail for the Lac-repressor headpiece, a protein with 51 amino acids consisting of 4 random coil sections and three alpha helices. In this work, the constraints were determined experimentally from NMR studies.

To demonstrate that our method is correct, we have applied PROTEAN to sperm whale myoglobin, a molecule whose crystal structure is known. By using distances between atoms in the crystal, distance constraints were applied to the eight helices in myoglobin to determine if PROTEAN would reproduce the crystal structure. The family of solutions obtained from PROTEAN includes the actual structure of the molecule. Work is proceeding to include the heme group of myoglobin as a component and use constraints to other portions of the molecule to further restrict the state families obtained.

D. Relevant Publications

1. Brinkley, J., Cornelius, C., Altman, R., Hayes-Roth, B. Lichtarge, O., Duncan, B., Buchanan, B.G., Jardetzky, O.: *Application of Constraint Satisfaction Techniques to the Determination of Protein Tertiary Structure.* Report KSL-86-28.

2. Buchanan, B.G., Altman, A., Brinkley, J., Cornelius, C., Duncan, B., Hayes-Roth, B., Hewett, M., Lichtarge, O., Jardetzky, O.: *The Heuristic Refinement Method for Deriving Solution Structures of Proteins.* Report KSL-85-41. October 1985. In *Proceedings of the National Academy of Sciences.*

3. Erman, L.D., Hayes-Roth, B., Lesser, V.R., Reddy, D.R.: *The HEARSAY-II Speech Understanding System: Integrating Knowledge to Resolve Uncertainty.* ACM Computing Surveys 12(2):213-254, June, 1980.

4. Hayes-Roth, B.: *The Blackboard Architecture: A General Framework for Problem Solving?* Report HPP-83-30, Department of Computer Science, Stanford University, 1983.

5. Hayes-Roth, B.: *BB1: An Environment for Building Blackboard Systems that Control, Explain, and Learn about their own Behavior.* Report HPP-84-16, Department of Computer Science, Stanford University, 1984.

6. Hayes-Roth, B.: *A Blackboard Architecture for Control.* Artificial Intelligence 26:251-321, 1985.

7. Hayes-Roth, B. and Hewett, M.: *Learning Control Heuristics in BB1.* Report HPP-85-2, Department of Computer Science, 1985.

8. Jardetzky, O.: *A Method for the Definition of the Solution Structure of Proteins from NMR and Other Physical Measurements: The LAC-Repressor Headpiece.* Proceedings of the International Conference on the Frontiers of Biochemistry and Molecular Biology, Alma Alta, June 17-24, 1984, October, 1984.

9. Hayes-Roth, B., Buchanan, B.G., Lichtarge, O., Hewitt, M., Altman, R., Brinkley, J., Cornelius, C., Duncan, B., and Jardetzky, O.: *PROTEAN: Deriving protein structure from constraints.* To appear in Proceedings of the AAAI, 1986.

E. Funding Support

Title: Interpretation of NMR Data from Proteins Using AI Methods

PI's: Oleg Jardetzky and Bruce G. Buchanan

Agency: National Science Foundation

Grant identification number: PCM 84-02348

Total Award Period and Amount: 11/1/84 - 10/31/86 $100,000
(includes direct and indirect costs)

Current award period and amount: 11/1/85 - 10/31/86 $ 50,000
(includes direct and indirect costs)

The following grants and contracts each provide partial funding for PROTEAN personnel.

Title: Research on Blackboard Problem-Solving Systems